Bank Loan Modelling: Predicting Personal Loan Acceptance

Author: Yash Makadia

Email: yashmakadia1908@gmail.com

Date: Oct 02, 2020

GitHub Repository: https://github.com/yash-makadia/Bank-Loan-Modelling

1. Introduction

The Bank Loan Modelling project aims to predict the likelihood of bank customers accepting personal loans, enabling targeted marketing campaigns to increase loan conversions while reducing costs. The bank seeks to convert liability-based customers (depositors) into asset customers (borrowers) to boost loan business revenue through interest. A previous campaign achieved a 9.6% conversion rate, and the bank now wants a data-driven model to identify high-probability customers for future campaigns.

This personal project leverages Python-based data science techniques, including exploratory data analysis (EDA), feature engineering, and classification modeling, to address this business problem. The primary objective is to build a model that maximizes the success ratio of loan acceptance predictions while minimizing marketing expenses.

2. Data Description

The dataset, sourced from Kaggle (https://www.kaggle.com/itsmesunil/bank-loan-modelling), contains data on 5,000 bank customers. It includes demographic, financial, and behavioral attributes, with the target variable indicating whether a customer accepted a personal loan in a previous campaign.

2.1 Dataset Attributes

- ID: Unique customer identifier (dropped during preprocessing).
- Age: Customer's age in years.
- Experience: Years of professional experience (dropped due to correlation with Age).
- Income: Annual income ($000).
- ZIP Code: Home address ZIP code (dropped as nominal).
- Family: Family size (1–4).
- CCAvg: Average monthly credit card spending ($000).
- Education: Education level (1: Undergrad, 2: Graduate, 3: Advanced/Professional).
- Mortgage: House mortgage value ($000).
- Personal Loan: Target variable (0: Did not accept, 1: Accepted; 9.6% acceptance rate).
- Securities Account: Has a securities account (0: No, 1: Yes).
- CD Account: Has a certificate of deposit account (0: No, 1: Yes).
- Online: Uses online banking (0: No, 1: Yes).
- CreditCard: Uses bank-issued credit card (0: No, 1: Yes).

2.2 Key Statistics

Size: 5,000 rows, 14 columns (initially).
Target Distribution: 480 customers (9.6%) accepted loans, indicating class imbalance.
Source: https://www.kaggle.com/itsmesunil/bank-loan-modelling

3. Methodology

The project follows a structured pipeline: data preprocessing, EDA, feature engineering, modeling, and evaluation.

3.1 Data Preprocessing

- **Loading Data**: Imported Bank_Personal_Loan_Modelling.xlsx using Pandas.
- **Cleaning**: Dropped ID (no predictive value), ZIP Code (467 unique values, nominal), and Experience (highly correlated with Age, contained negative values: -1, -2, -3 for 52 rows). Verified no missing values using df.isnull().sum().
- **Data Types**: Confirmed appropriate types (e.g., Income: int64, CCAvg: float64, Personal Loan: int64).

3.2 Exploratory Data Analysis (EDA)

EDA was conducted to understand feature distributions and relationships with the target variable.

3.2.1 Univariate Analysis

- Age: Normally distributed (mean: 45.34 years, std: 11.46).
- Income: Right-skewed (mean: $73.77K, max: $224K).
- CCAvg: Right-skewed (mean: $1.94K, max: $10K).
- Mortgage: Highly skewed (69.24% have zero mortgage; max: $635K).
- Family: Balanced distribution (1: 29.4%, 2: 25.9%, 3: 20.2%, 4: 24.4%).
- Education: 41.9% Undergrad, 28.1% Graduate, 30% Advanced.
- Securities Account: 10.4% have accounts (522 customers).
- CD Account: 6% have accounts (302 customers).
- Online: 59.7% use online banking.
- CreditCard: 29.4% use bank-issued credit cards.
- Personal Loan: 9.6% acceptance rate (480 customers).

3.2.2 Bivariate Analysis

- Income vs. Education: Customers with Undergrad education (level 1) have higher incomes, but loan acceptance is similar across education levels.
- CD Account vs. Personal Loan: Nearly all customers with CD Accounts accepted loans, a strong predictor.
- Family vs. Personal Loan: Family size 3 is slightly more likely to accept loans.
- Securities Account: Most non-loan customers have securities accounts, suggesting a negative correlation.
- Correlation: Income and CCAvg are highly correlated (visualized via heatmap).

3.2.3 Visualizations

Histograms for continuous variables (Age, Income, CCAvg, Mortgage).
Count plots for categorical variables (Family, Education, CreditCard, Online).
Box plots (Income vs. Education by Personal Loan).
Count plots with hue (Securities Account, CD Account, Family by Personal Loan).
Correlation heatmap and pair plots for feature relationships.
Pie chart showing 9.6% loan acceptance rate.

3.3 Feature Engineering

Transformations: Applied Yeo-Johnson transformation to Income and CCAvg to reduce skewness, improving logistic regression performance. Binned Mortgage into discrete intervals (0–100, 100–200, ..., 600–700) to handle skewness and zero-inflation (69.24% zeros).

Standardization: Used StandardScaler to normalize features for model compatibility.

Feature Selection: Dropped ID, ZIP Code, and Experience based on EDA insights.

3.4 Modeling

Data Split: Split data into 70% training and 30% testing sets with stratification to maintain class balance.

Models:

1. **Logistic Regression:** Baseline model with default parameters.
2. **Random Forest Classifier:** Ensemble model with 500 estimators and max depth 8.
3. **Decision Tree Classifier:** Non-parametric model with max depth 8.
4. **Naive Bayes:** Gaussian Naive Bayes for probabilistic classification.
5. **Evaluation Metrics:** Accuracy, precision, recall, F1-score, ROC AUC score. Confusion matrices to assess true positives (loan acceptances) and false positives.

3.5 Libraries

- pandas, numpy: Data manipulation.
- matplotlib, seaborn: Visualization.
- scikit-learn: Preprocessing, modeling, and evaluation.
- jupyter: Interactive notebook environment.

4. Results and Discussion

Random Forest outperformed all models with 98.73% accuracy and 0.94 ROC AUC.

4.1 Model Performance

The following table summarizes the performance of the four models on the test set (1,500 samples):

Model	Accuracy (%)	F1-Score (Class 1)	ROC AUC	Confusion Matrix (TN, FP, FN, TP)
Logistic Regression	95.47	0.73	0.82	[1338, 18, 50, 94]
Random Forest	98.73	0.93	0.94	[1353, 3, 16, 128]
Decision Tree	98.00	0.89	0.93	[1344, 12, 18, 126]
Naive Bayes	91.53	0.55	0.75	[1294, 62, 65, 79]

- Random Forest outperformed others with 98.73% accuracy, an F1-score of 0.93 for the positive class (loan acceptance), and a ROC AUC of 0.94, due to its ability to handle imbalanced data and complex feature interactions.
- Logistic Regression achieved 95.47% accuracy but had a lower F1-score (0.73) and ROC AUC (0.82), indicating weaker performance on the minority class.
- Decision Tree performed well (98% accuracy, F1-score 0.89), but was slightly less robust than Random Forest.
- Naive Bayes had the lowest performance (91.53% accuracy, F1-score 0.55), likely due to its assumption of feature independence.

4.2 Key Findings

Feature Importance: CD Account is a strong predictor; nearly all customers with CD Accounts accepted loans. Higher Income and CCAvg correlate with loan acceptance. Family size 3 shows a slight tendency to accept loans. Education level 1 (Undergrad) customers have higher incomes, but loan acceptance is consistent across levels.

1. **Data Imbalance**: The 9.6% loan acceptance rate required careful handling (stratified splitting, F1-score focus).
2. **Skewness**: Transforming Income, CCAvg, and Mortgage improved model performance, especially for Logistic Regression.
3. **Visualizations**: Heatmaps, box plots, and count plots revealed critical relationships (e.g., CD Account vs. Personal Loan).

4.3 Visualizations

1. **Univariate**: Histograms showed Age as normally distributed, while Income, CCAvg, and Mortgage were right-skewed (addressed via transformations).
2. **Bivariate**: Box plots highlighted income differences across education levels; count plots showed CD Account holders' high loan acceptance.
3. **Correlation**: Heatmap confirmed high correlation between Income and CCAvg.
4. **Model Evaluation:** Confusion matrix heatmaps visualized true positives (e.g., 128 for Random Forest) and false positives (e.g., 3 for Random Forest).
5. **Loan Distribution:** Pie chart illustrated the 9.6% acceptance rate.

5. Business Understanding

The bank aims to increase personal loan conversions by targeting liability customers with a higher likelihood of acceptance. The Random Forest model, with 98.73% accuracy and a 0.93 F1-score, is recommended for deployment due to its superior performance. Key business recommendations include:

1. **Targeted Marketing:** Prioritize customers with CD Accounts, as they are highly likely to accept loans. Focus on customers with higher incomes and credit card spending. Consider family sizes of 3, which show a slight propensity for loans.
2. **Campaign Optimization:** Use the Random Forest model to score customers by loan acceptance probability, enabling targeted campaigns to high-probability customers. Minimize false positives (e.g., only 3 in Random Forest) to reduce marketing costs on unlikely candidates.
3. **Business Impact:** Increase conversion rates beyond the previous 9.6% by focusing on high-probability customers. Reduce campaign costs by targeting fewer, more promising customers. Enhance customer satisfaction by offering tailored loan products to likely acceptors.

6. Conclusion

The Bank Loan Modelling project successfully developed a predictive model to identify bank customers likely to accept personal loans. The Random Forest Classifier, with 98.73% accuracy and a 0.94 ROC AUC, emerged as the best model, driven by strong predictors like CD Account, Income, and CCAvg. The project addressed data challenges (imbalance, skewness) through preprocessing and feature engineering, delivering actionable insights for the bank's marketing team.

6.1 Future Scope

1. Incorporate additional data (e.g., transaction history) for richer features.
2. Explore advanced models like XGBoost or Neural Networks.
3. Develop a real-time prediction API for marketing integration.
4. Create an interactive dashboard to visualize customer segments and predictions.

7. References

Kaggle Dataset: https://www.kaggle.com/itsmesunil/bank-loan-modelling

Python Libraries:

Pandas: https://pandas.pydata.org/

NumPy: https://numpy.org/

Matplotlib: https://matplotlib.org/

Seaborn: https://seaborn.pydata.org/

Scikit-learn: https://scikit-learn.org/

Contact: For questions or feedback, please contact Yash Makadia at yashmakadia1908@gmail.com.